Biais par la taille

Djalil Chafaï[1]

Hiver 2013

Problème. *Pour tout entier $k \geq 1$, soit p_k la fréquence des foyers de taille k dans la population française, de sorte que $\sum_{k\geq 1} p_k = 1$. Prenons à présent un français au hasard dans la population. La taille T du foyer auquel il appartient est aléatoire. Montrer que $\mathbb{P}(T=k) \approx \frac{k}{m} p_k$ pour tout $k \geq 1$, où par définition $m := \sum_{k\geq 1} k p_k$ est la taille moyenne des foyers.*

La loi de T n'est pas la loi de départ, car les grands foyers sont sur-représentés tandis que les petits foyers sont sous-représentés. Ce phénomène est appelé *biais par la taille* (*size-bias* en anglais). Il s'agit sans doute du biais d'échantillonnage le plus connu. Le biais est d'autant plus important que la taille du foyer diffère de la taille moyenne m.

Solution. On se place conditionnellement à la population, qui compte au total $\sum_{k\geq 1} N_k$ foyers où N_k désigne le nombre de foyers de taille k. La population compte au total $N_1 + 2N_2 + \cdots$ individus. On modélise le choix d'un français au hasard par le tirage d'un entier selon la loi uniforme sur l'intervalle $[\![1, N_1 + 2N_2 + \cdots]\!]$. Il y a N_k foyers de taille k qui comptent au total kN_k individus. Par définition de la loi uniforme (formule « cas favorables sur cas totaux ») la probabilité que ce français appartiennent à un foyer de taille k est

$$\frac{kN_k}{N_1 + 2N_2 + \cdots} = \frac{k}{m} p_k \quad \text{où} \quad p_k := \frac{N_k}{N_1 + N_2 + \cdots} \quad \text{et} \quad m := p_1 + 2p_2 + \cdots .$$

□

Variante. Partant des fréquences $p_1, p_2, \ldots$, on peut aussi choisir de modéliser la taille des foyers français par une suite $X_1, \ldots, X_n$ de v.a. i.i.d. de loi $\mathbb{P}(X_1 = k) = p_k$ pour tout $k \geq 1$, et de moyenne $m := \mathbb{E}(X_1) = \sum_{k\geq 1} p_k$. La France compte beaucoup de foyers, et donc n est grand (millions !). La population compte au total $X_1 + \cdots + X_n$ individus. On modélise le choix d'un français au hasard par le tirage, conditionnellement à $X_1, \ldots, X_n$ (c'est-à-dire sachant $X_1, \ldots, X_n$), d'un entier selon la loi uniforme sur l'intervalle $[\![1, X_1 + \cdots + X_n]\!]$. Pour tout $k \geq 1$, il y a $N_k := \mathbf{1}_{\{X_1=k\}} + \cdots + \mathbf{1}_{\{X_n=k\}}$ foyers de taille k qui comptent au total kN_k individus. Par conséquent, par définition de la loi uniforme (formule « cas favorables sur cas totaux ») et à une double application de la loi forte des grands nombres :

$$\mathbb{P}(T_n = k | X_1, \ldots, X_n) = \frac{k \sum_{i=1}^n \mathbf{1}_{\{X_i=k\}}}{\sum_{i=1}^n X_i} = \frac{k \frac{1}{n} \sum_{i=1}^n \mathbf{1}_{\{X_i=k\}}}{\frac{1}{n} \sum_{i=1}^n X_i} \xrightarrow[n\to\infty]{\text{p.s.}} \frac{k}{m} p_k.$$

À présent, par convergence dominée, on obtient enfin le résultat souhaité :

$$\mathbb{P}(T_n = k) = \mathbb{E}(\mathbb{P}(T_n = k | X_1, \ldots, X_n)) \underset{n\to\infty}{\longrightarrow} \frac{k}{m} p_k.$$

Combinatoire. Il est possible d'étudier le biais par la taille avec une modélisation purement combinatoire préfréquentiste. Considérons le tirage de $n \leq N$ boules dans une urne contenant $N = N_1 + \cdots + N_d$ boules dont N_k boules de couleur k pour tout $1 \leq k \leq d$. La formule « cas favorables sur cas totaux » pour le modèle d'équiprobabilité dit que pour tous $n_1 \leq N_1, \ldots, n_d \leq N_d$, la probabilité d'obtenir n_k boules de couleur k pour tout $1 \leq k \leq d$ vaut

$$\frac{\binom{N_1}{n_1} \cdots \binom{N_d}{n_d}}{\binom{N}{n_1,\ldots,n_k}} = \frac{\frac{N_1!}{n_1!(N_1-n_1)!} \cdots \frac{N_d!}{n_d!(N_d-n_d)!}}{\frac{N!}{n_1!\cdots n_d!}} = \frac{N_1! \cdots N_d!}{N!(N_1-n_1)! \cdots (N_d-n_d)!}.$$

1. http://djalil.chafai.net/

Il s'agit de la loi hypergéométrique $\mathrm{HypGeo}((N_1,\dots,N_d))$. Lorsque $\min(N_1,\dots,N_d)\to\infty$ avec $N_k/N\to p_k$ pour tout $1\le k\le d$ alors $p_1+\cdots+p_d=1$ et $\mathrm{HypGeo}((N_1,\dots,N_d))$ converge vers la loi multinomiale $\mathrm{Multinom}(n,(p_1,\dots,p_d))$ de taille n et de paramètre $(p_1,\dots,p_d)$ (formule de Stirling !). Maintenant, on décide que d est la taille maximale d'un foyer dans la population, et que N est le nombre de foyers de la population. À présent, dans notre urne, on remplace chaque boule de couleur k par k (petites) boules de couleur k. La nouvelle urne contient alors

$$M:=\sum_{k=1}^{d}kN_k$$

boules, dont kN_k boules de couleur k pour tout $1\le k\le d$. On effectue à présent un tirage de n boules dans cette nouvelle urne. La formule « cas favorables sur cas totaux » pour le modèle d'équiprobabilité dit que pour tous $(n_1,\dots,n_d)$ tels que $n_k\le kN_k$ pour tout $1\le k\le d$, la probabilité d'obtenir n_k boules de couleur k pour tout $1\le k\le d$ vaut

$$\frac{\prod_{k=1}^{d}(kN_k)!}{M!\prod_{k=1}^{d}(kN_k-n_k)!}.$$

Il s'agit du biais par la taille de la loi $\mathrm{HypGeo}((N_1,\dots,N_d))$. Lorsque $\min(N_1,\dots,N_d)\to\infty$ avec $N_k/N\to p_k$ pour tout $1\le k\le d$, alors cette loi converge vers la loi multinomiale de taille n et de paramètre $(1p_1,\dots,dp_d)$ (qui est le biais par la taille de la loi $(p_1,\dots,p_d)$). Notons que $M/N=\sum_{k=1}^{d}kN_k/N\to\sum_{k=1}^{d}kp_k=m$.

Poisson. La notion de biais par la taille reste bien entendu valable si $p_0\neq 0$. Plus précisément, si $p_0,p_1,\dots$ est une loi sur $\mathbb{N}$ de moyenne $m=\sum_{k\ge0}kp_k=\sum_{k\ge1}kp_k$ alors son biais par la taille est la loi $p_1/m,p_2/m,\dots$ sur $\mathbb{N}^*$. Pour les lois de Poisson, le biais par la taille se traduit par une translation à droite : si $Z\sim\mathrm{Poi}(\lambda)$ alors son biais par la taille a la loi de $1+Z$.

Identifiabilité, débiaisage, estimation. Si $p_0,p_1,\dots$ et $p'_0,p'_1,\dots$ sont deux lois sur $\mathbb{N}$ de moyennes m et m' et telles que $p_0=p'_0$ et $\frac{k}{m}p_k=\frac{k}{m'}p'_k$ pour tout $k\ge1$ alors en divisant par k puis en sommant sur k il vient $m=m'$ et enfin $p=p'$. Si $q_1,q_2,\dots$ est une loi sur $\mathbb{N}^*$ et si $0<\alpha\le1$ alors $p_0:=1-\alpha,p_1:=\alpha q_1,p_2:=\alpha q_2,\dots$ est une loi sur $\mathbb{N}$ dont le biais par la taille ne dépend pas de α. Il faut donc connaître p_0 pour pouvoir identifier p. Revenons au cas où $p_0=0$ pour simplifier. D'un point de vue statistique, pour estimer les p_k à partir d'un échantillon $Y_1,\dots,Y_n$ de loi $q_1,q_2,\dots$ où $q_k:=\frac{k}{m}p_k$ pour tout $k\ge1$, on peut estimer q_k par $\hat{q}_k:=\frac{1}{n}\mathrm{card}\{1\le i\le n:Y_i=k\}$ (note : $\sum_k\hat{q}_k=n$) puis m par $\hat{m}:=\frac{1}{\sum_k\frac{1}{k}\hat{q}_k}$ puis p_k par $\hat{p}_k:=\frac{\hat{m}}{k}\hat{q}_k$. Peut-on faire mieux ?

— oOo —